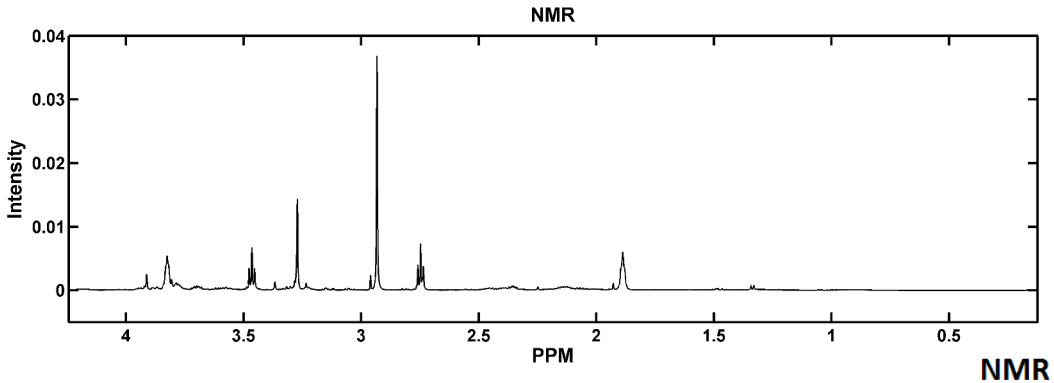
Using Galaxy for Metabolomics

ICG8
2013
Rob L Davidson PhD

Overview

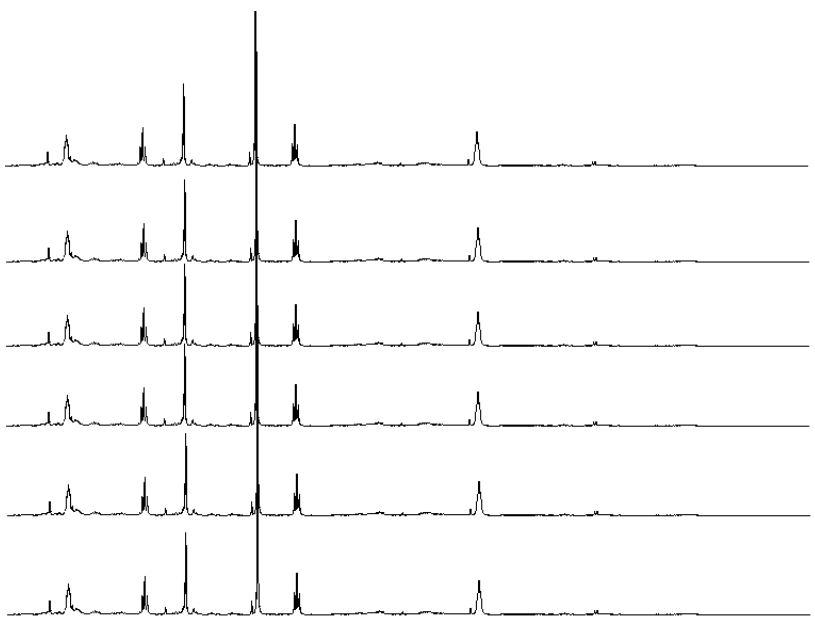
- Metabolomics
- Galaxy
- Birmingham Workflow
- Galaxy Implementation

Metabolomics - data



NMR spectrum = vector

	Peak 1	Peak 2	Peak3	Peak 4...	Peak N
Sample 1	19812	432	2309	4501882	5876



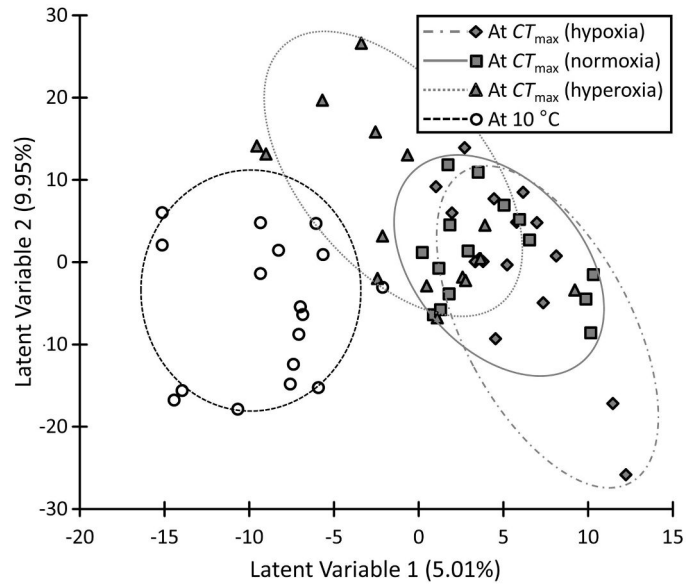
NMR spectra = matrix

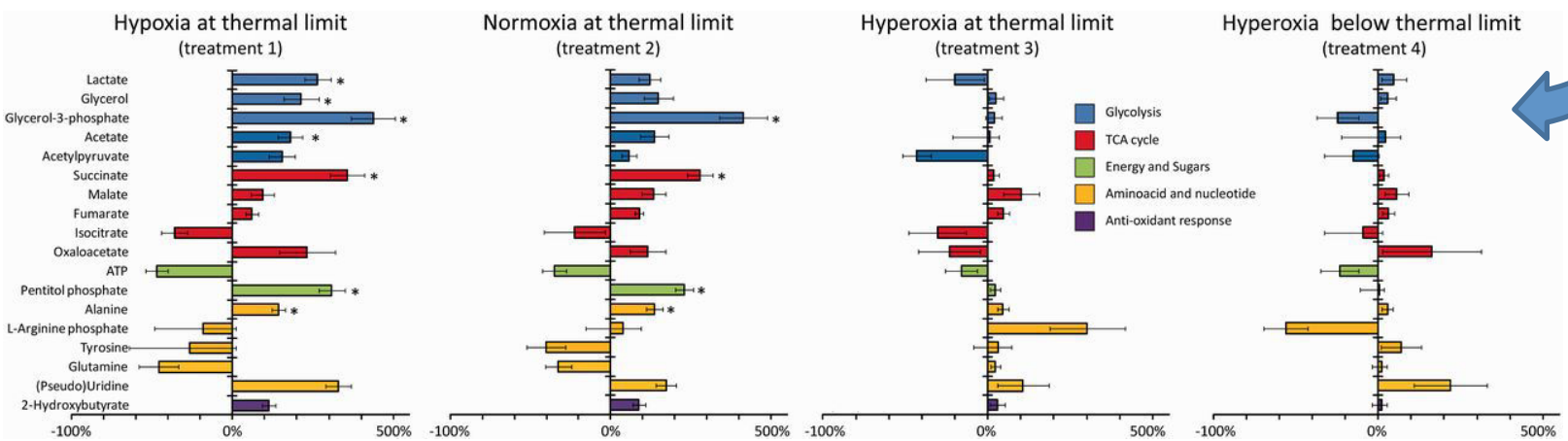
	Peak 1	Peak 2	Peak3	Peak 4...	Peak N
Sample 1	19812	432	2309	4501882	5876
Sample 2	8994	654	5409	357890	312
Sample 3	15012	1098	3102	1342098	10879
Sample 4	9999	302	4231	1809282	890
Sample N	17531	789	4500	2200192	3456

Metabolomics - tools

NMR spectra = matrix

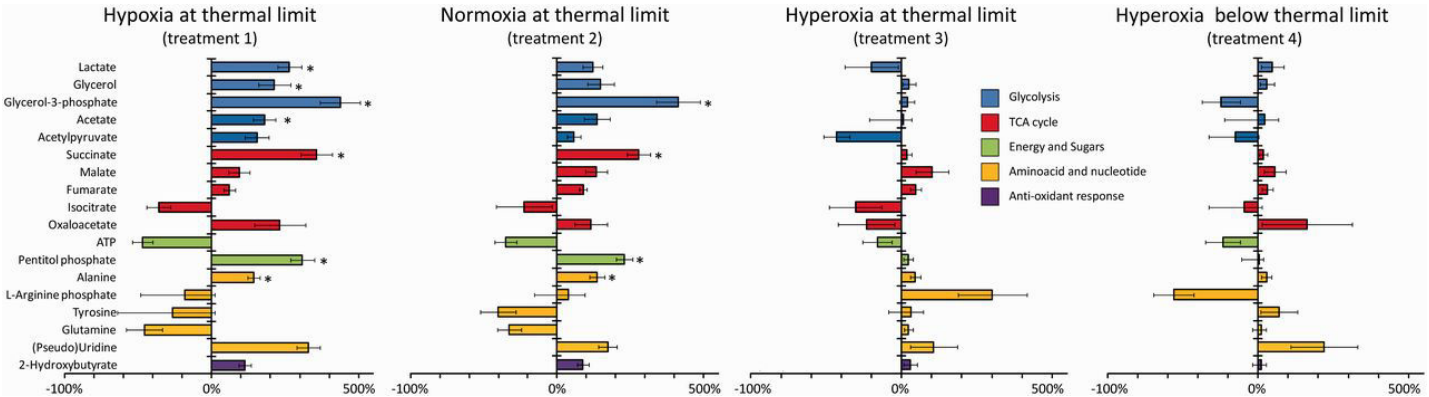
	Peak 1	Peak 2	Peak3	Peak 4...	Peak N
Sample 1	19812	432	2309	4501882	5876
Sample 2	8994	654	5409	357890	312
Sample 3	15012	1098	3102	1342098	10879
Sample 4	9999	302	4231	1809282	890
Sample N	17531	789	4500	2200192	3456

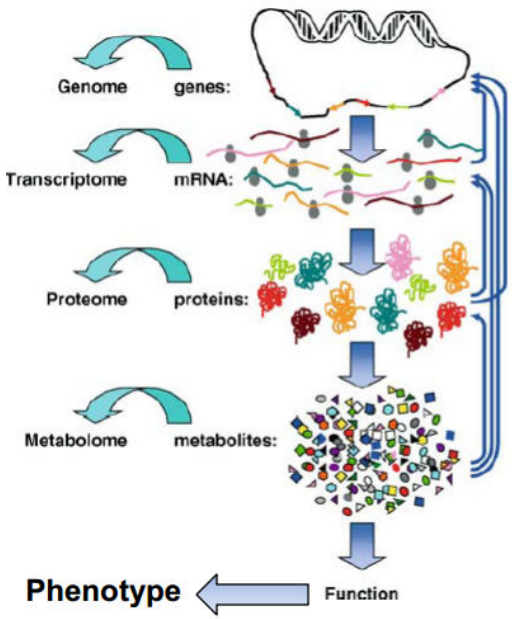




Verberk WCEP, Int&Comp Biol. 2013

Metabolomics - future



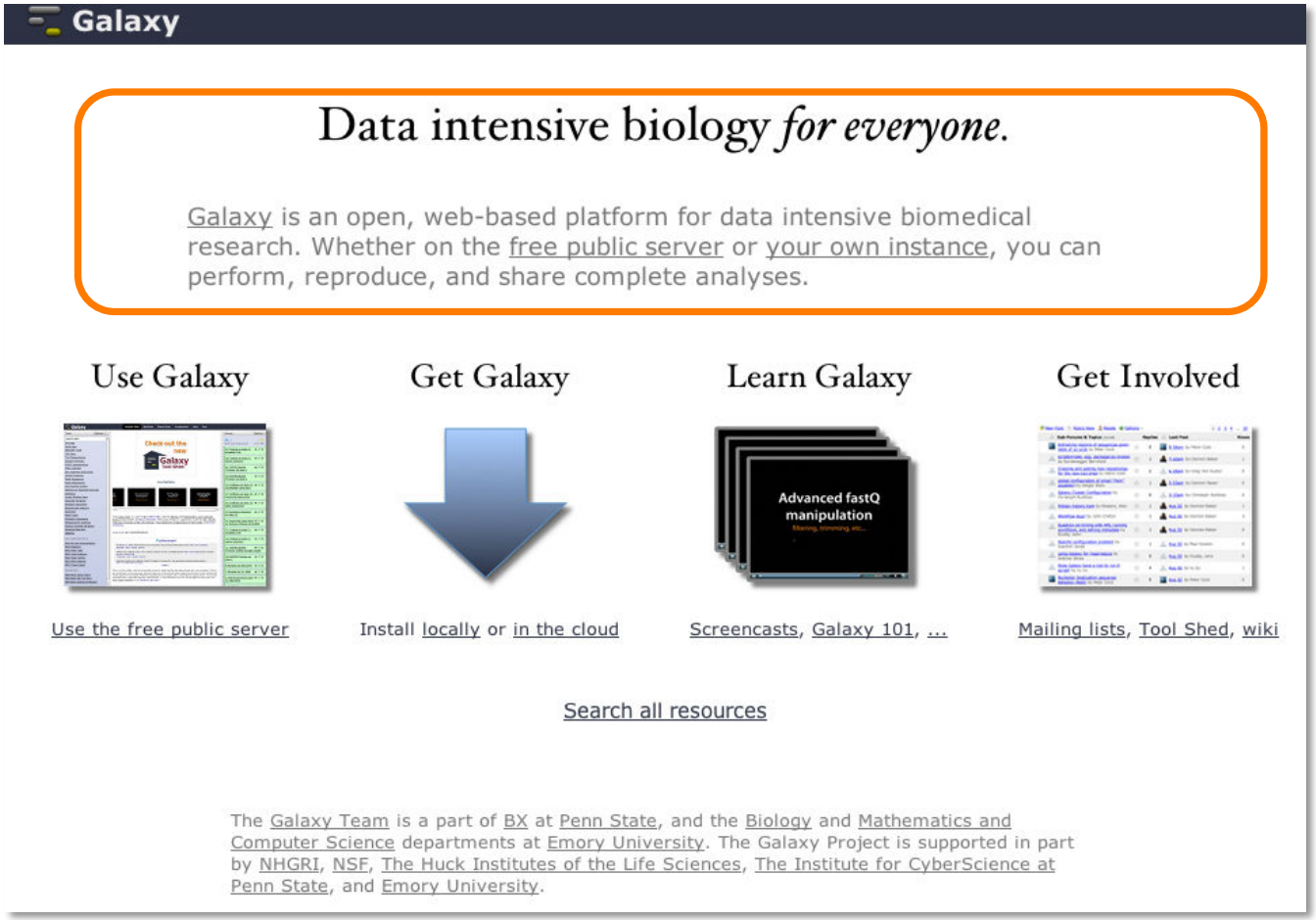


- Identify metabolite profile
- Link to pathways etc
- Correlate with other ‘omes
- **Trans-omics!**

Goodacre, R. Metabolomics (2005)

Galaxy

Galaxy



Open source

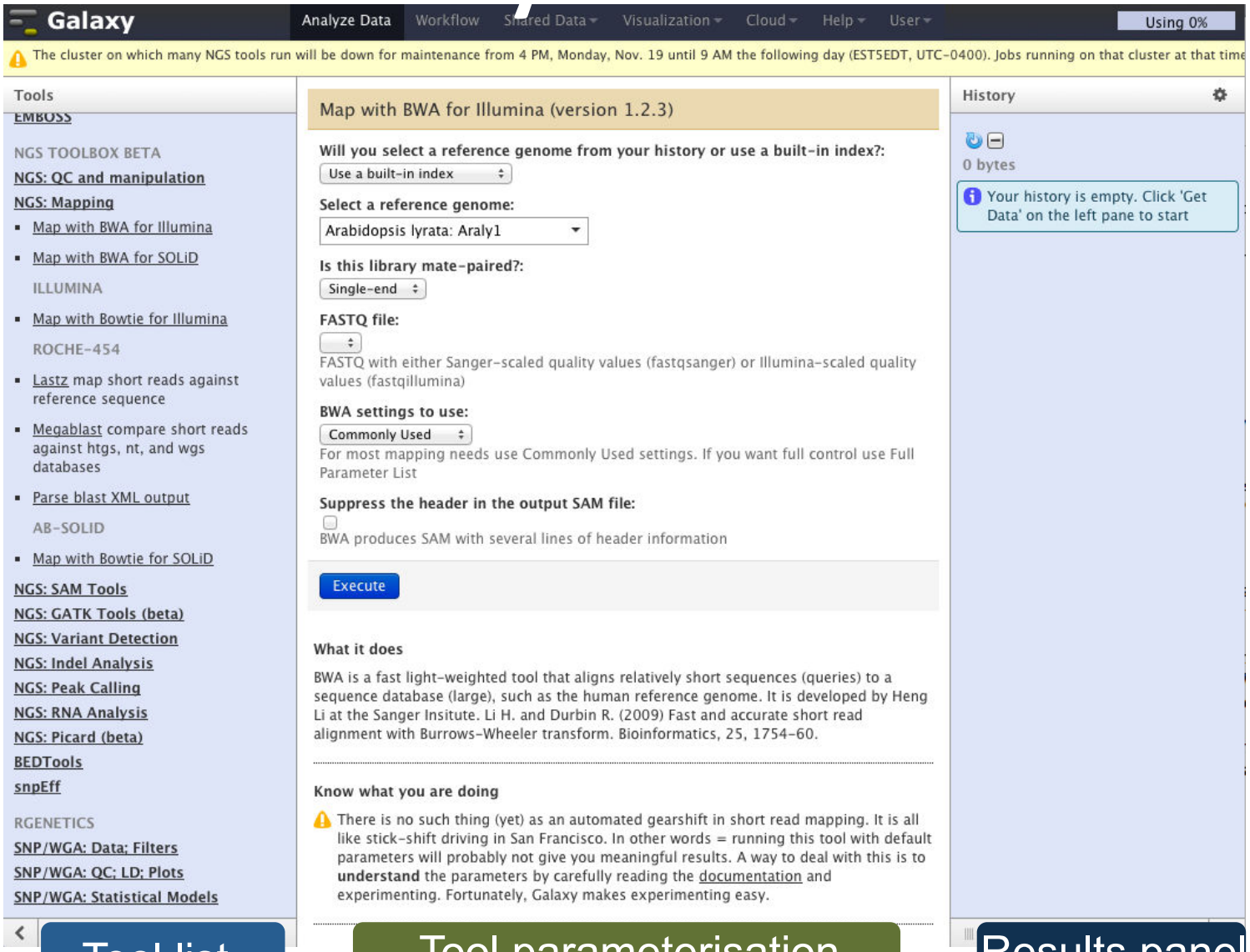
Over 20,000 main Galaxy server users

Over 1200 papers citing Galaxy use

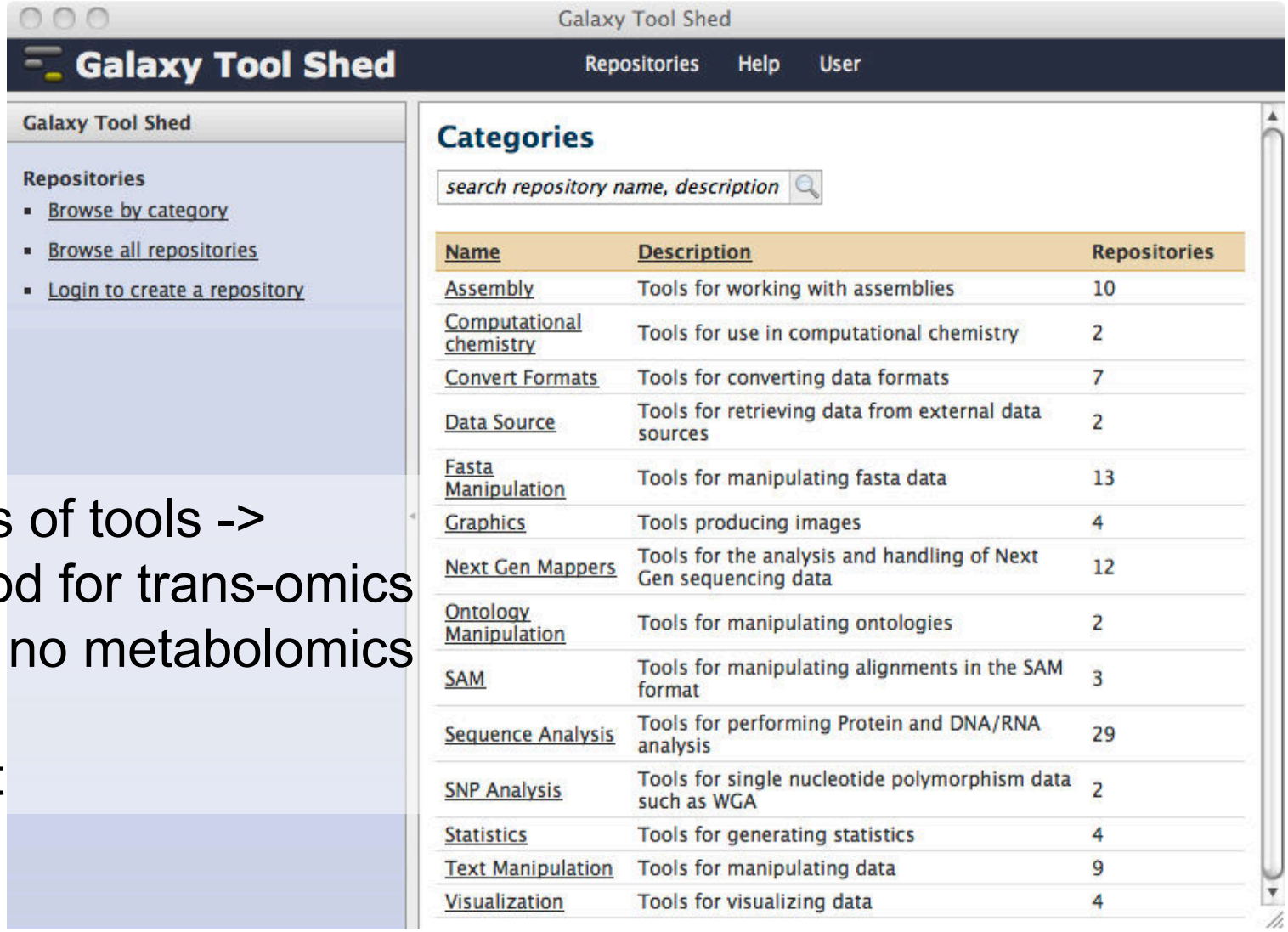
45+ public Galaxy servers

http://galaxyproject.org

Galaxy - interface



Galaxy - toolshed



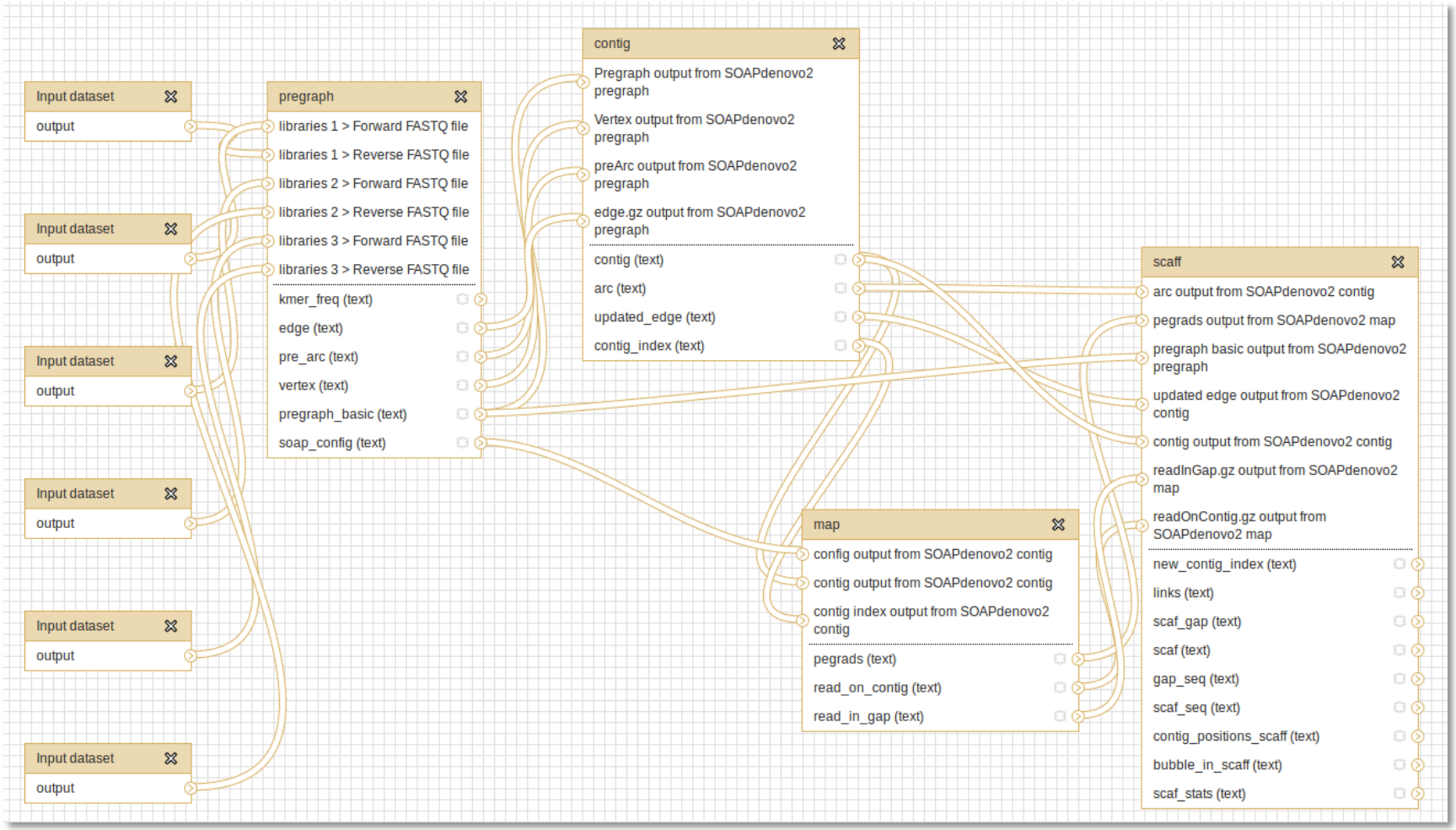
Name	Description	Repositories
Assembly	Tools for working with assemblies	10
Computational chemistry	Tools for use in computational chemistry	2
Convert Formats	Tools for converting data formats	7
Data Source	Tools for retrieving data from external data sources	2
Fasta Manipulation	Tools for manipulating fasta data	13
Graphics	Tools producing images	4
Next Gen Mappers	Tools for the analysis and handling of Next Gen sequencing data	12
Ontology Manipulation	Tools for manipulating ontologies	2
SAM	Tools for manipulating alignments in the SAM format	3
Sequence Analysis	Tools for performing Protein and DNA/RNA analysis	29
SNP Analysis	Tools for single nucleotide polymorphism data such as WGA	2
Statistics	Tools for generating statistics	4
Text Manipulation	Tools for manipulating data	9
Visualization	Tools for visualizing data	4

- Lots of tools ->
- Good for trans-omics
- But no metabolomics

…yet

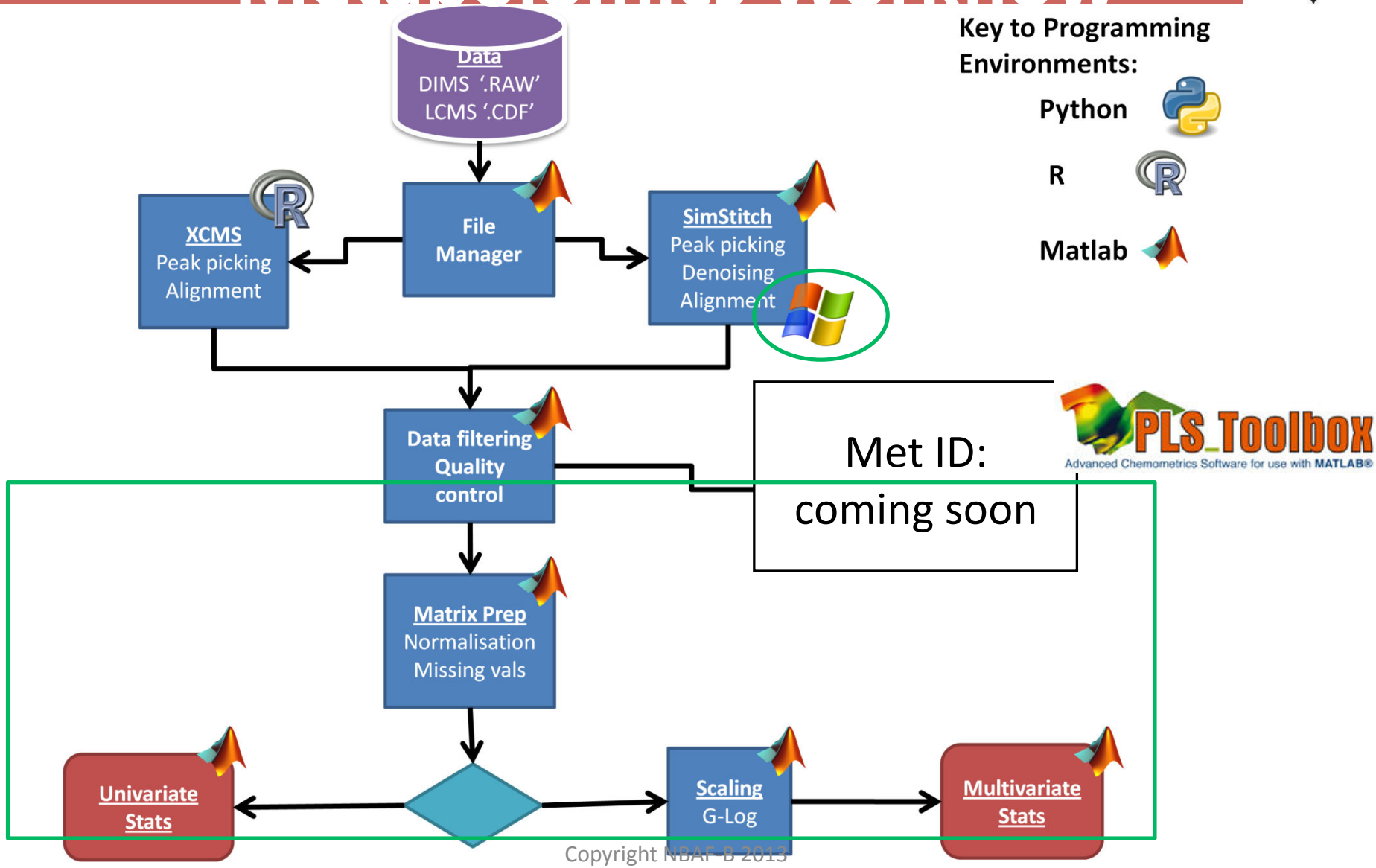
http://toolshed.g2.bx.psu.edu/

Galaxy - workflows

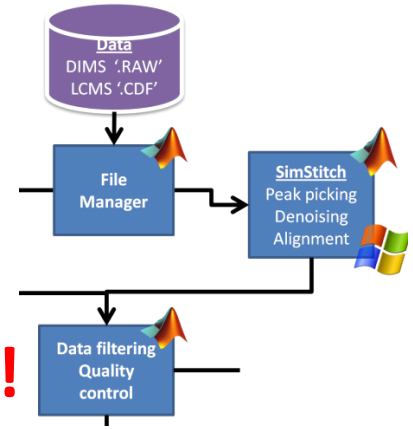


Metabolomics Workflow

Metabolomics Workflow

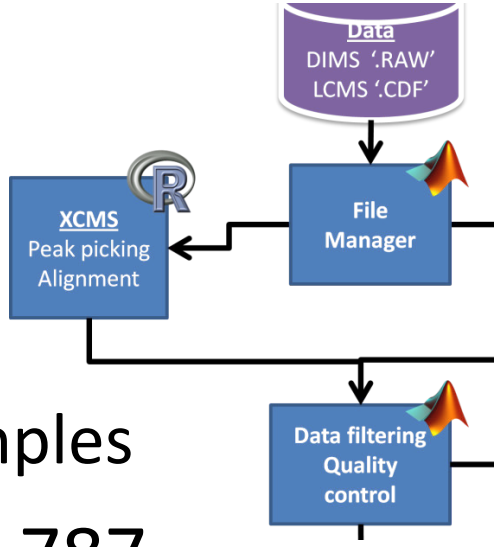


Metabolomics Workflow



- **SimStitch**
  - FTICR-DIMS data
  - RAW format ; **NB: requires MS Windows!**
  - Stitches together short MZ ranges for greater accuracy
  - Picks peaks/removes noise, aligns samples
  - Applies filters to technical replicates, blanks and samples for greater robustness
- *Southam AD, Anal Chem. 2007;79(12):4595-602*.

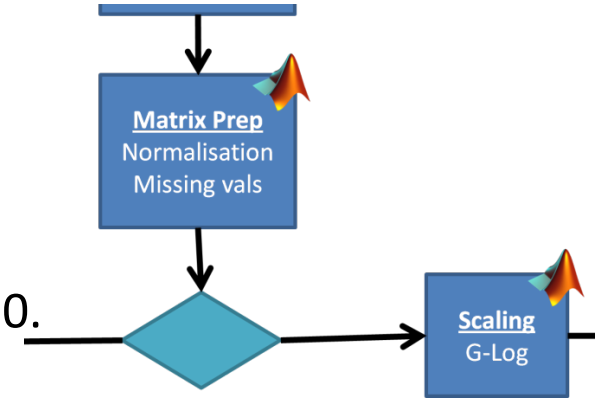
Metabolomics Workflow



- **XCMS**
  - LC-MS (also does GC-MS etc)
  - netCDF format (also does MZML etc)
  - Picks peaks/removes noise, aligns samples
- Smith CA, Anal. Chem. 2006 78:779-787
- http://www.bioconductor.org/packages/2.12/bioc/html/xcms.html

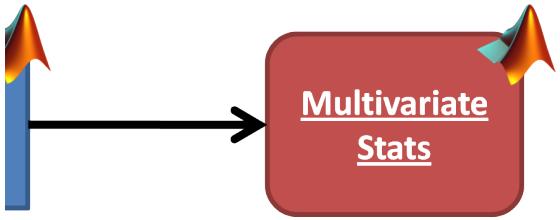
- In our pipeline, we call XCMS (in R) using Matlab...
  - To make use of our FileManager Structure
  - Because we also use Matlab for post-XCMS processing

Metabolomics Workflow



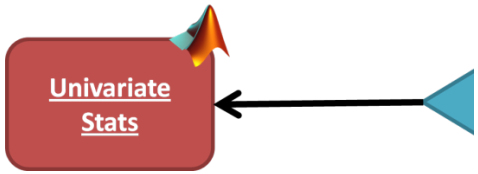
- **Matrix Prep:**
  - PQN Normalisation
    - Dieterle F, Anal Chem. 2006;78(13):4281-90.
  - KNN-Missing Value Imputation
    - Hrydziuszko O, Metabolomics 2012 8(1):161-174
  - G-Log scaling and variance stabilisation:
    - Parsons HM,BMC Bioinf. 2007 8:234
  - All done using PLS Toolbox data structures in Matlab

Metabolomics Workflow



- **Multivariate stats**
  - PCA with automatic selection of PCs
  - 2 classes: T-Test on scores on each PC
  - 3+ classes: ANOVA and Tukey-Kramer
  - **Output = text file containing these statistics**
  - a PLS Toolbox 'model' is also created and scores plots etc can be viewed in Matlab

Metabolomics Workflow



- **Univariate stats**
  - 2 classes: T-Test or Mann-Whitney-U for each peak
  - 3+ classes: ANOVA or Kruskall-Wallis
  - False Discovery Rate correction (Benjamini Hochberg)
  - **Output = csv file containing these statistics**

Metabolomics Workflow

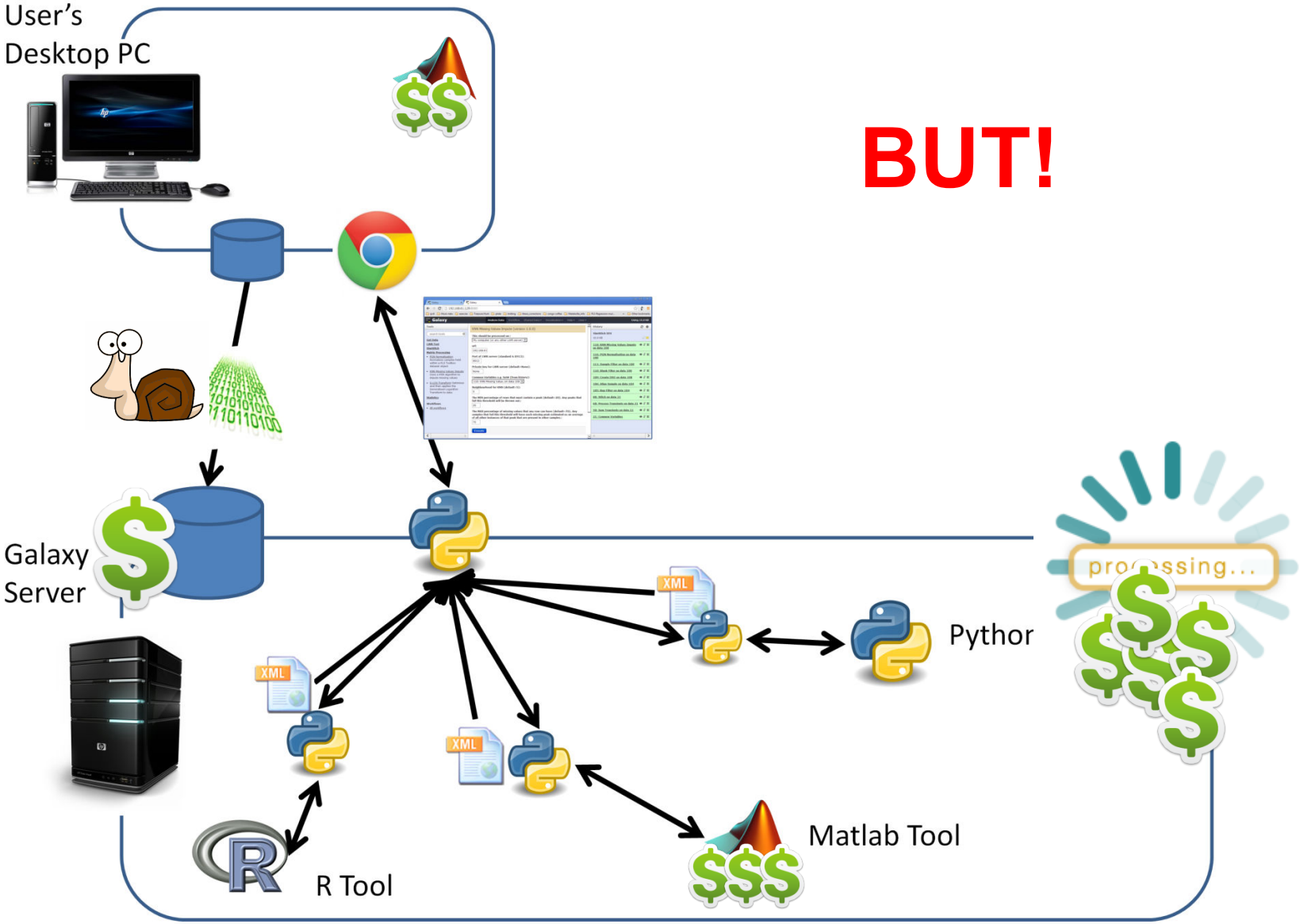
- **Done**
  - 1st end-end metabolomics pipeline in Galaxy
  - FTICR-DIMS and LCMS data
- **To Do**
  - Add in MI-Pack (underway)
  - Add more stats e.g. PLSDA
    - possibly merging with Netherlands Metabolomics http://galaxy.nmcdsp.org/ (stats only)
  - Replace input file structure with ISA-Tab
    - http://www.ebi.ac.uk/metabolights/

Metabolomics Workflow

- **Requires 32bit MS Windows**
- **Data is large (100s Gb per study)**
- **Lots of processing power**
- **Multiple licenses**

Galaxy Implementation

Standard Galaxy



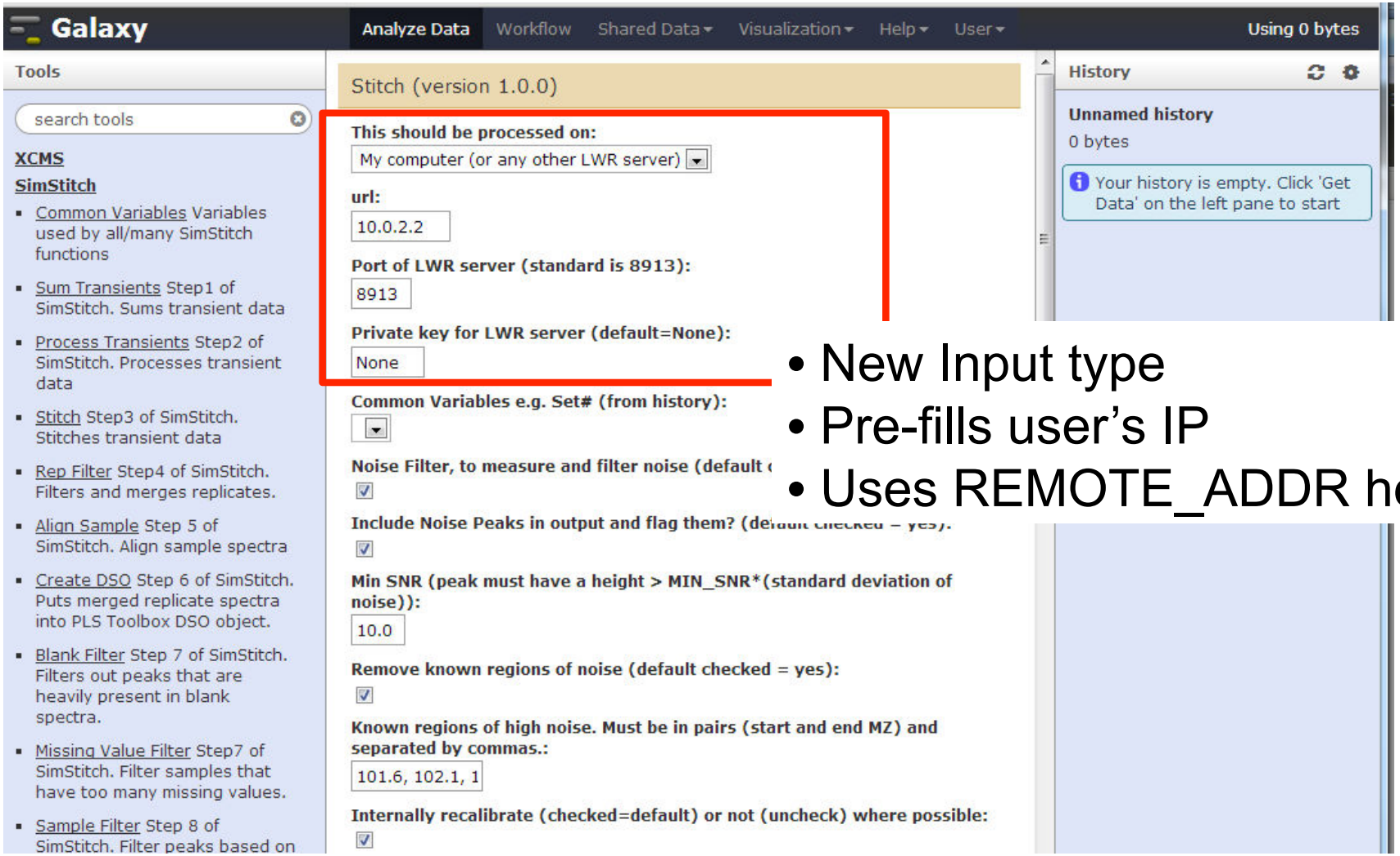
Metabo - Galaxy

- Requirements
  - Allow Galaxy access to MS Windows
    - for FTICR-DIMS RAW file processing
  - Avoid passing large data over LAN
    - slow
  - Minimise cost of Galaxy implementation
    - make use of existing processors, storage and licences

Metabo - Galaxy

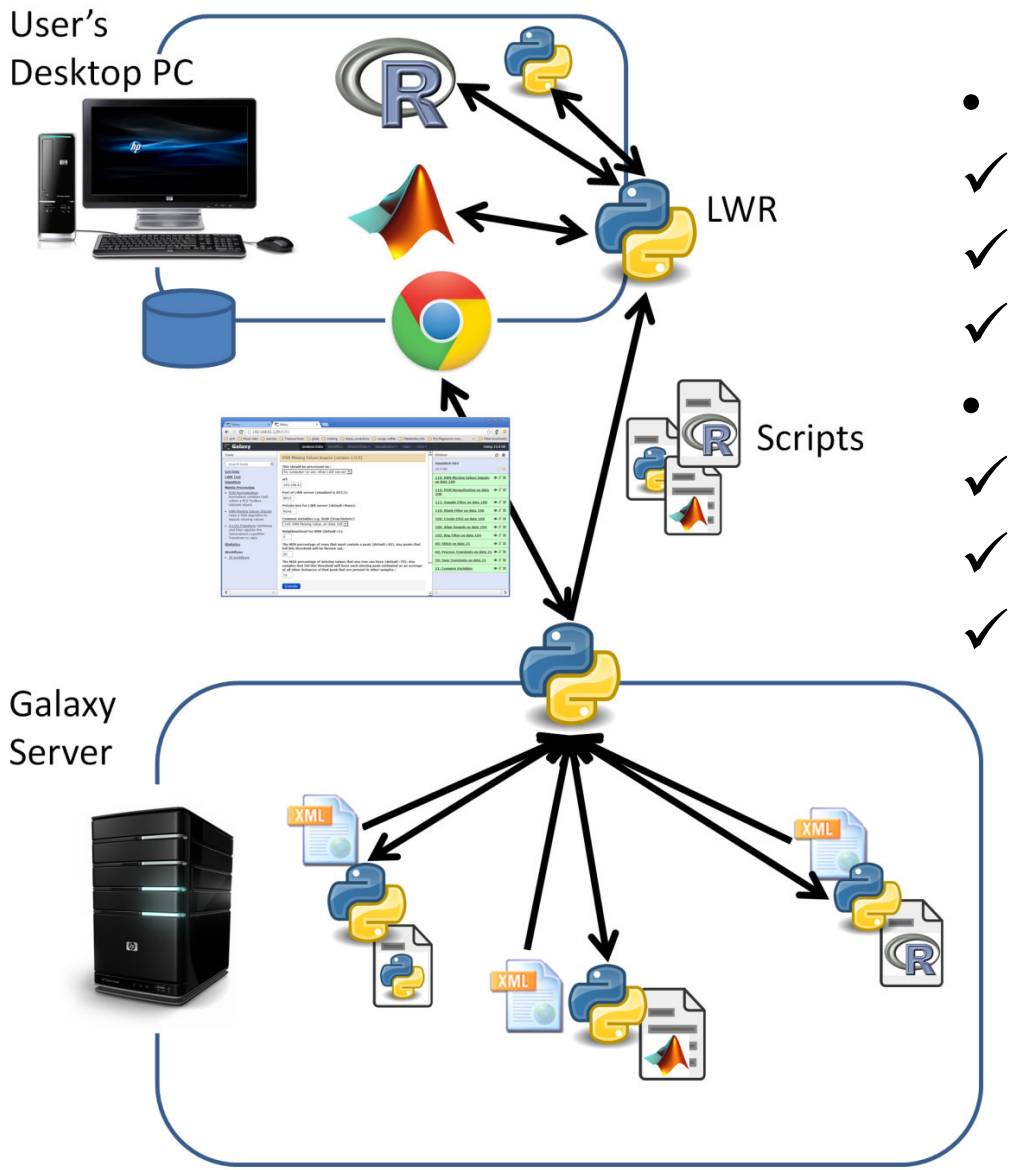
- Solution
  - Use Galaxy's Light Weight Runner (LWR)
  - Install LWR client on user's desktop
  - Adjust Python wrappers to send tools via LWR
  - Run all tools on User's desktop (MS Windows)
  - No need for
    - extra licenses
    - central storage/file transfer
    - powerful server

Metabo - Galaxy



- New Input type
- Pre-fills user's IP
- Uses REMOTE_ADDR header

Metabo - Galaxy



- **Makes use of**
- ✓ existing proc. power
- ✓ licenses
- ✓ user's MS Windows (RAW)
- **Still acts as**
- ✓ version control
- ✓ workflow manager
- ✓ GUI

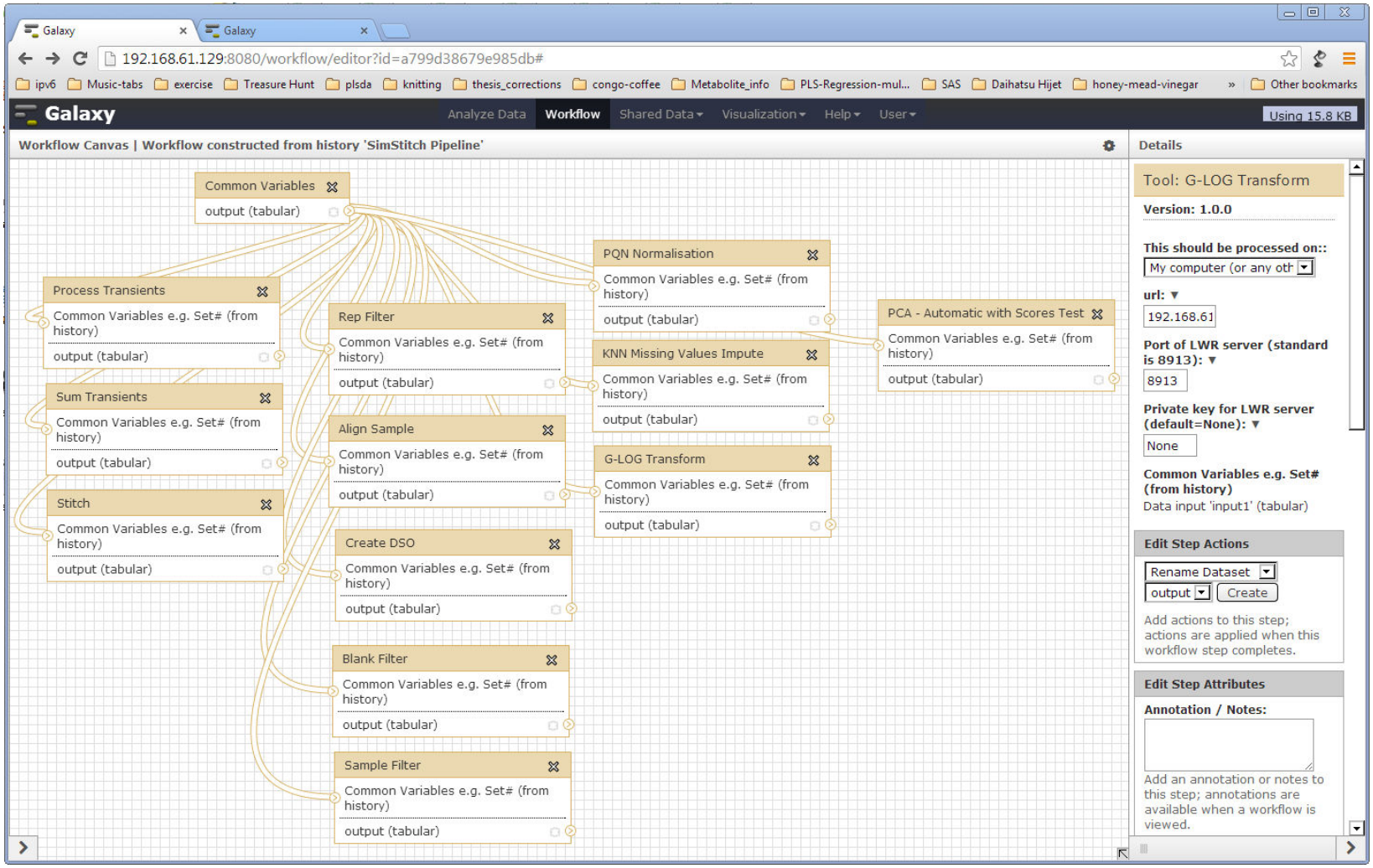
Metabo - Galaxy

That said...

- Working with GigaScience
  - http://www.gigasciencejournal.com/
- To be hosted on GigaGalaxy
  - http://galaxy.cbiit.cuhk.edu.hk/
- Using normal setup
  - (all processing/licenses on Galaxy server)
- Downloadable version to include both options

Summary

First RAW -> stats Galaxy Pipe



Summary

- Metabolomics has entered Galaxy!
- Can be expanded BY COMMUNITY!

- Can merge more easily with other ‘omics

- Have developed Galaxy in a new way that allows
  - Use of existing hardware
  - Use of existing licenses
  - Less slow transfer of large data

Acknowledgements

- University of Birmingham
  - Ralf Weber, Ulf Sommer, Mark Viant
- Gigascience
  - Pete Li
- NERC Discipline Hop scheme

End

Questions?

r.l.davidson@bham.ac.uk